퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 1, pp. 124-129

Ranking Translation Word Selection Using a Bilingual Dictionary and WordNet

Kweon Yang Kim* and Se Young Park**

* School of Computer Engineering, Kyungil University

** Dept. of Computer Engineering, Kyungpook National University

Abstract

This paper presents a method of ranking translation word selection for Korean verbs based on lexical knowledge contained in a bilingual Korean-English dictionary and WordNet that are easily obtainable knowledge resources. We focus on deciding which translation of the target word is the most appropriate using the measure of semantic relatedness through the 45 extended relations between possible translations of target word and some indicative clue words that play a role of predicate-arguments in source language text. In order to reduce the weight of application of possibly unwanted senses, we rank the possible word senses for each translation word by measuring semantic similarity between the translation word and its near synonyms. We report an average accuracy of 51% with ten Korean ambiguous verbs. The evaluation suggests that our approach outperforms the default baseline performance and previous works.

Key words : translation word selection, semantic similarity, near synonym.

1. Introduction

In the context of machine translation, picking the correct translation for a target word among multiple candidates is known as translation word selection. The resolution of sense ambiguity and selection of correct translation in non-restricted text are perhaps the great central problem at the lexical level of machine translation.

In recent, corpus based solutions for translation word selection have become quite popular. However, theses approaches heavily rely on the availability of manually created sense tagged or bilingual parallel corpora that are expensive to create. An alternative approach is to take advantage of the knowledge available in machine readable dictionary such as WordNet[4].

In this paper, we present a method of ranking translation words that attempts to select the most appropriate translation word for Korean verbs using the measure of semantic relatedness that the WordNet similarity package[10] provides. Knowledge sources are extracted from a bilingual Korean-English dictionary and WordNet. From Korean-English dictionary, we first extract the possible English translation words that are corresponding to each sense of given Korean words.

The solution suggested in this paper is to identify the semantic relatedness between possible translation words in the view of target language context. Therefore, when investigating senses of possible translation words for a given source language word, we do not know which sense is more appropriate for each translation word. Considering all possible senses for each translation word introduces the application of unwanted senses in the view of source language word that may bring about an incorrect translation word selection. In order to avoid the application of unwanted senses, we rank the possible word senses for each translation word by measuring semantic similarity in WordNet between the translation word and near synonyms that are described in a bilingual dictionary.

From WordNet, we exploit the knowledge for which the semantic relatedness between the senses of translation words pair for source language words pair is measured.

접수일자 : 2006년 1월 4일
완료일자 : 2006년 2월 13일
감사의 글 : This research was supported by Kyungpook National University Research Fund, 2005

2. Previous Works

Approaches based on statistical machine translation exploit bilingual parallel corpora or hand-crafted translations for selecting translation words[2]. However, as Koehn and Knight[7] pointed out, those knowledge sources are generally hard to acquire.

Dagan and Itai[3] have proposed an approach for resolving word sense ambiguity and selecting a translation word using statistical data from a monolingual corpus of target language. They use the statistics on lexical co-occurrence within syntactic relations in a monolingual corpus of target language. Their method can relieve the knowledge acquisition problem by relying on only a monolingual corpus of target language that is sense untagged. However, their method is apt to select an in-

appropriate translation word because it is based on the simple mapping information between a source word and its possible translation words by using only a bilingual dictionary and a sense untagged target language corpus. When seeing an occurrence of target language word, their method counts word occurrence as a translation word of all the source language words. This can mislead the selection of correct translation word because the translation word also may have several senses from the view of target language.

The source of problem is that this method considers the unnecessary senses of translation words that do not have anything to do with the senses of the original source language word. A possible solution is to use the sense tagged monolingual corpora or bilingual parallel corpora[7, 9]. However, knowledge sources like sense tagged or parallel corpora are generally hard to come by because they need a manual creation. Consider the translation of Korean phrase "nonmun-ul ssuda"(that means *write paper*). Noun 'nonmun' has only one sense and has its corresponding translation word *paper* in Korean-English dictionary. For simplicity's sake, we took into consideration only the first two senses of the translation word *paper* such as

paper#n#1
{paper}
-- (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)

paper#n#2
{composition, **paper**, report, *theme*}
-- (an essay (especially one written as an assignment);
"he got an A on his composition")

In addition, the Korean verb 'ssuda' has several translation words such as *write, put on, use, employ, speak,* and etc that are corresponding equivalent to each respective sense in Korean-English dictionary. Table 1 shows the list of combination pairs for possible translation words between 'nonmun' and 'ssuda'.

For this example, the method of Dagan and Itai finds the pairs of possible translation words for source language words using a bilingual dictionary and estimates probabilities of how likely each combination pair of translation words occurs in a target language corpus. We find that an incorrect translation word use for a given target word 'ssuda'is selected because the co-occurrence frequency of a pair of translation word paper and use is the highest in a target language corpus.

For above example, their method fails to select correct translation word for the target word 'ssuda'. This is because the sense of translation word *paper* in almost all of occurrences of *use* and *paper* in target language corpus is used as a first sense of *paper*(paper#n#1 : the first sense of paper as a noun) in WordNet. In consequence, this method selects incorrect translation words pair *use-paper* instead of correct translation words pair *write-paper.*

Table 1 Combination pair of translation words with its co-occurrence frequency in the target language corpus

source words pair	translation words pair		counts
	Verb	Noun	
'nonmun' : Noun	*write*	*paper*	22
	put on	*paper*	0
'ssuda' : Verb	***use***	***paper***	**47**
	employ	*paper*	0
	speak	*paper*	0
	...		

3. Dictionary based Word Sense Disambiguation

Word sense disambiguation is the task of assigning to each occurrence of an ambiguous word in a text one of its possible senses. In recent, corpus based solutions for word sense disambiguation have become popular. In general, such approaches rely on the availability of sense tagged corpus that needs a manual creation. An alternative is dictionary based approaches that take advantage of the information available in machine readable dictionaries or lexical database.

The Lesk algorithm[8] may be identified as a starting point for dictionary based approaches. His algorithm does not require any sense tagged training instances. This approach disambiguates word senses for a particular target word by comparing the dictionary glosses of its possible senses with those of its surrounding context words. The target word is assigned the sense whose gloss has the most overlapping words with the glosses of its surrounding words. The intuition of this approach is based on two facts that the sense of target word is semantically related to the surrounding words in context and the overlaps of sense glosses reflect the degree of semantic relatedness between senses of a target word and senses of surrounding words in context. Such connections are indicative of the senses in which these words are used rather than not by chance.

The main problem of Lesk algorithm is that dictionary glosses tend to be quite short and may not provide sufficient information to measure of relatedness between a given target word and surrounding words in context. To overcome the limitation of Lesk algorithm based on short definitions, recently Banerjee and Pedersen[1] presented an extended method of semantic relatedness measurement through the concept hierarchy of WordNet rather than simply considering the glosses of the surrounding words.

They suggest an extended approach that exploits not

only explicit relations through direct links but also implicit relations through indirect links that WordNet provides. In their experiments, they used the seven possible relations such as synset, hypernym, hyponym, holonym, meronym, troponym, and attribute of each word sense in WordNet. In effect, the glosses of surrounding words in the context are expanded to include glosses of theses related words to which they are semantically related through the extended relations in WordNet.

This measurement is not restricted by any specific part of speech and can be applied for measurement of relatedness between senses from any part of speech. All of relations that WordNet provides are not equally helpful and the optimal selection of the best relations depends on the application of overlap measurement.

4. Ranking Translation Word Selection

Our algorithm for ranking translation words is based on the hypothesis that the correct translation of a target word is highly related to senses of translation words for surrounding context words in the view of target language. The Korean-English dictionary and WordNet provide good information for mapping from the source language word to possible translation words and mapping from the translation word to word senses. The algorithm first identifies a set of possible translation words for the target word and surrounding context words using a Korean-English bilingual dictionary.

Figure 1 shows part of the Korean-English dictionary. Such a bilingual dictionary generally classifies the meaning of an entry word into several senses and groups its translation words and near translation synonyms by each sense class.

기사[1]/gisa/ 技師
an engineer; a technician
기사[2]/gisa/ 記事
an article; an account; news; a statement; description; election news ... , a stoppress ...
기사[3]/gisa/ 騎士
a rider; a horseman

논문[1]/nonmun/ 論文
a paper; a treatise; a dissertation; a thesis; write a paper.

Fig. 1 A part of the Korean-English dictionary

In Korean-English dictionary, each Korean lexical entry word may have multiple senses and there are some corresponding translation words for each sense. In Figure 1, a Korean noun 'gisa' has three senses and has three representative translation words *engineer, article,* and *rider* with near translation synonyms *technician, account, news, statement, description* and *horseman* according to each sense. In the case of Korean noun 'nonmun' entry, there is only one sense and has a representative translation word paper with near synonyms such as *treatise, dissertation,* and *thesis.*

Near translation synonyms[5] are words that have close senses with the representative translation word for each sense. In addition to sense ambiguity of source language word, the translation word may have multiple senses in WordNet. For example, the translation word *paper* that is corresponding equivalent for source word 'nonmun' has two different noun senses such as paper#n#1 and paper#n#2 as described in section 2.

In order to reduce the weight of application of possibly unwanted senses for the translation word that may bring about incorrect translation word selection, our algorithm needs to decide which sense of the representative translation word is more appropriate for the source language word before we start to compute the score of semantic relatedness between combination pairs of possible senses. The weight of m-th sense t_k^{lm} for translation word t_k^l is computed as follows:

$$w(t_k^{lm}) = \frac{1}{rank\ (t_k^{lm})}$$

The rank of sense t_k^{lm} is computed using the semantic similarity score between the possible senses of the representative translation word and near translation synonyms of t_k^l. The measure of semantic similarity between two senses is computed by using the method of Jiang and Conrath which combines the information contents of each sense and their least common subsumer that returns the largest set of commonalities between each sense[6]. For the word *paper* that is the translation of a source word 'nonmun', the sense paper#n#2 has higher weight than the sense paper#n#1 because the sense paper#n#2 is more semantically similar to near synonyms of translation word paper such as *treatise, dissertation,* and *thesis.*

After applying the weight of each sense, our algorithm computes the score of semantic relatedness between combination pairs of possible senses. This algorithm selects the translation word of target word that has maximum relatedness using the following equation.

$$\arg\max_{i=1}^{x1} \sum_{k \in clues} \max_{j,l,m=1}^{x2,x3,x4} rel(t_0^{ij}, t_k^{lm}) \bullet w(t_0^{ij}) \bullet w(t_k^{lm})$$

This equation computes a relatedness score of each translation word t_0^i for a target word t_0. t_k^{lm} means m-th sense of translation word t_k^l of t_k that is a surrounding clue word.

Two senses are considered to be related semantically when they share a set of words that occur in their respective glosses. The counting of number of overlaps between just gloss words is not likely to find relatedness

because they are too short to result in a reliable measure. However, if we consider the glosses and usage examples of all the senses that are connected directly and indirectly to a sense by extended relations, this becomes a larger set of words that can measure relatedness. Rather than simply considering the glosses of the surrounding words in the context, the concept hierarchy of WordNet is exploited to allow for glosses of word senses related to the context words to be compared.

WordNet is a lexical database widely used in natural language processing. The wide availability of WordNet has led to the development of a number of approaches to word sense disambiguation problem. Each sense of a word is mapped to a synset in WordNet and all synsets in WordNet are linked by a variety of semantic relations. In addition to glosses associated with each sense, WordNet provides many example usages associated with them that play a role of sense tagged training instances. These usage examples and glosses are not likely to be much help because they are short. However, the number of example usages and glosses grow larger by extending through the set of relations that WordNet provides.

In our experiments, we adopt the following extended relations that WordNet provides. Figure 2 shows the 45 extended relations between noun and verb we adopt.

```
/* direct(noun)-direct(verb) */
syns-syns; syns-glos; syns-exam
glos-syns ; glos-glos; glos-exam
exam-syns; exam-glos; exam-exam

/* coordinate(noun)-direct(verb) */
syns(coor)-syns; syns(coor)-glos; syns(coor)-exam
glos(coor)-syns; glos(coor)-glos; glos(coor)-exam
exam(coor)-syns; exam(coor)-glos; exam(coor)-exam

/* direct(noun)-coordinate(verb) */
syns-syns(coor); glos-syns(coor); exam-syns(coor)
syns-glos(coor); glos-glos(coor); exam-glos(coor)
syns-exam(coor); glos-exam(coor); exam-exam(coor)

/* hyponyms(noun)-direct(verb) */
syns(hypo)-syns; syns(hypo)-glos; syns(hypo)-exam
glos(hypo)-syns; glos(hypo)-glos; glos(hypo)-exam
exam(hypo)-syns; exam(hypo)-glos; exam(hypo)-exam

/* direct(noun)-troponyms(verb) */
syns-syns(trop); glos-syns(trop); exam-syns(trop)
syns-glos(trop); glos-glos(trop); exam-glos(trop)
syns-exam(trop); glos-exam(trop); exam-exam(trop)
```

Fig. 2 Extended relations (syns : synsets, glos : definitions, exam : examples, coor : coordinate, hypo : hyponyms, trop : troponyms)

While Pedersen and Banerjee used the immediately surrounding context words, we choose the most relevant words that provide much more indicative clues for selecting correct translation word. The focus in much recent work is on local collocation knowledge including a variety of distance and syntactic relations. For the English tasks, that local collocation knowledge such as n-grams is expected to provide important evidences to resolve sense ambiguity because English is a fixed order language. However, Korean is a partially free order language and therefore the ordering information on surrounding words of the ambiguous word does not provide significantly meaning information for resolving sense ambiguity in Korean.

Korean has some particularities: plenty of inflectional verb endings, postpositions instead of preposition, and so on. The postpositions and verb endings represent syntactic relations such as predicate-arguments, modifier-modified relations. We deal three major predicate-arguments relations: verb-object, verb-locative and verb-instrument relation that construct complement-head structure. The postpositions attached to the noun are either 'ul'/'rul', 'ae', or 'ro'/'uro' that usually represent syntactic relations between the noun and the verb: object, location, and instrument respectively.

In this paper, we focus the selection of translation word for Korean transitive verbs. Therefore we decided to adopt the three major predicate-argument relations such as Noun-'ul'/'rul' (object) + Verb, Noun-'ae' (locative) + Verb, and Noun-'ro'/'uro' (instrument) + Verb rather than immediate surrounding words to improve the performance of translation word selection.

In order to extract syntactic features, we need a robust syntactic parser. However, the accuracy performance of current parser is not high. Therefore we have taken a simple partial parser by using an only part of speech tagger without robust parsing to identify syntactic relations such as "the first noun with postposition 'ul'/'rul', 'ae', or 'ro'/'uro'to the left or right of target word in context". This provides a better approximation for syntactic relations than just requiring a limited distance between surrounding words in the context, while relying on the availabilityof part of speech tagger that is simpler and more available than robust syntactic parser. Sometimes it is hard to identify the such syntactic relation in the case of embedded sentence. For example, a sentence "gang-i bumramha-nun gut-ul makda/*keep the river from overflowing*" has the first noun 'gut-ul'(ING) for a verb 'makda'. In this case, the algorithm extracts the nominal form 'bumram'/overflow of preceding verb 'bumramha'/overflow as a verb-object relation.

5. Experiments

In order to evaluate our algorithm for ranking translation words, 1.4 million words Gemong Korean encyclo-

pedia is used as a test data set. Our experiments are performed on ten Korean transitive verbs, 'makda', 'ssuda', 'seuda', 'japda', 'mandulda', 'nanuda', 'mutda', 'batda', 'utda', and 'jitda' that appear with high frequency and ambiguity in our experiment corpus.

This data set consists of 12,513 instances each of which contains a sentence with a target verb to be translated. In order to compare our result with the human created answer, human experts have annotated these instances manually with correct translation words in advance.

All the results we report are given as the accuracy performance which means the number of correct translations divided by the number of answers. The results of experiment are summarized in Table 2. The performance ofour algorithm was compared with the baseline performance(Baseline) that is achieved by assigning all occurrences to the most frequent translation word. The experimental results show that the accuracy performance of our algorithm(RTW: Ranking Translation Word) performs better than the default baseline(Baseline), original Lesk algorithm(Lesk) and Pedersen and Banerjee's(P&B).

Table 2 Experimental results

Target Verbs	# of Senses	Accuracy(%)			
		Base line	Lesk	P&B	RTW
makda	7	27	18	38	52
ssuda	9	21	15	41	62
seuda	10	40	22	40	48
japda	8	32	20	37	45
mandulda	10	29	18	42	48
nanuda	6	45	17	48	64
mutda	7	35	18	25	45
batda	19	38	21	35	49
utda	11	30	17	40	56
jitda	13	28	15	36	45
Average	10	33	18	38	51

6. Conclusions

In this paper we presented a method of ranking translation word selection based on a bilingual dictionary and WordNet that are easily obtainable knowledge sources. The translation words are selected by measuringsemantic relatedness through the several extended relations between possible translations of target word and surrounding clue words in source language text. Considering all possible senses for each translation word introduces the application of unwanted senses in the view of source language word that may bring about an incorrect translation word selection. In order to avoid the application of unwanted senses, we first rank the possible word senses for each translation word by measuring semantic similarity in WordNet between the translation word and near synonyms that are described in a bilingual dictionary.

The evaluation suggests that our approach performs better than other approaches including the baseline performance. We report an average accuracy of 51% with ten Korean ambiguous verbs. While the results are generally lower than the supervised approaches, these results are significant because our approach is based on a bilingual dictionary and a target language dictionary, therefore it provides an alternative to the corpus based approaches that need the supervised learning.

References

[1] Banerjee, S. and Pedersen, T., "Extended Gloss Overlaps as a Measure of Semantic Relatedness," Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 805-810, 2003.

[2] Brown, P., Cocke, J., Pietra, V., Peitra, S., Jelinek, F., Lafferty, J., Mercer, R. and Roosin, P., "A Statistical Approach to Machine Translation, Computational Linguistics," Vol. 16, No. 2, pp. 79-85, 1990.

[3] Dagan, I. And Itai, A., "Word Sense Disambiguation Using a Second Language Monolingual Corpus," Computational Linguistics, Vol. 20, No. 4, pp. 563-596, 1994.

[4] Fellbaum, C., WordNet: An Electronic Lexical Database, MIT Press, 1998.

[5] Inkpen, D. and Hirst, G., "Automatic Sense Disambiguation of the Near-Synonyms in a Dictionary Entry," Proceedings, Fourth Conference on Intelligent Text Processing and Computational Linguistics, pp. 258-267, 2003.

[6] Jiang, J. and Conrath, D., "Semantic Similarity based on Corpus Statistics and Lexicon Taxonomy," Proceedings on International Conference on Research in Computational Linguistics, pp. 19-33, 1997.

[7] Koehn, P. and Knight K., "Knowledge Sources for Word-Level Translation Models," Empirical Methods in Natural Language Processing Conference, pp. 27-35, 2001.

[8] Lesk, M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: how to tell a pine code from an ice cream cone," Proceedings of the Fifth Annual International Conference on Systems Documentations, pp. 24-26, 1986.

[9] Li, H. and Li, C., "Word Translation Disambiguation

Using Bilingual Boostrapping," Computational Linguistics, Vol. 30, No. 1, pp. 1-22, 2004.
[10] Patwardhan, S. and Perdersen, T., The cpan wordnet::similarity package, http://search.cpan.org/~sid/WordNet-Similarity-0.06/, 2005.

저 자 소 개

김권양(Kweon Yang Kim)
1983년 : 경북대학교 전자공학과(학사)
1990년 : 경북대학교 전자공학과(석사)
1998년 : 경북대학교 컴퓨터공학과(박사)
1983년~1988년 : ETRI 연구원
1999년~2000년 : University of Central Florida 방문교수
1991년~현재 : 경일대학교 컴퓨터공학부 교수

관심분야 : 한글공학, 시멘틱웹
Phone : 053-850-7287
Fax : 053-850-7609
E-mail : kykim@kiu.ac.kr

박세영(Se Young Park)
1980년 : 경북대학교 전자공학과(학사)
1982년 : KAIST 전산학과(석사)
1989년 : 프랑스 파리 제7대학(박사)
1982년~2000년 : ETRI 책임연구원
2003년~2005년 : IITA 전문위원
2005년~현재 : 경북대학교 컴퓨터공학과 교수

관심분야 : 정보검색, 시멘틱웹, 디지털 콘텐츠
Phone : 053-950-5550
Fax : 053-959-4846
E-mail : seyoung@mail.knu.ac.kr